[redacted]

IR NO. 69-40

INFORMAL REPORT

A FREE FIELD RAPID ACCESS FORMAT FOR BT PROFILES

MARCH 1969

NAVAL OCEANOGRAPHIC OFFICE
WASHINGTON, D. C. 20390

INFORMAL REPORT

The Informal Report (IR) as produced at the Naval Oceanographic Office is a means for personnel to issue timely scientific and technical preliminary reports of their investigations. These are primarily informal documents used to report preliminary findings or useful byproducts of investigations and work to members of the scientific and industrial communities.

Informal Reports are assigned sequential numbers for each calendar year; the digits preceding the dash indicate the year.

The distribution made of this report is determined primarily by the author. Information concerning obtaining additional copies or being placed on a distribution list for all future Informal Reports in a given area of interest or specialty field, should be obtained from:

Distribution Control Department
Code 4420
Naval Oceanographic Office
Washington, D.C. 20390

ABSTRACT

A format of the restructured bathythermograph file is specified. The methods used to achieve a file memory requirement reduction ratio of about 9 to 1 are described.

WALTER E. YERGEN

Ocean Dynamics Branch
Exploratory Oceanography Division
Research and Development Department

APPROVED FOR RELEASE: R. L. Winokur
Acting Director
Exploratory Oceanography Division

DATE: 21 March 1969

TABLE OF CONTENTS

FIGURES

INTRODUCTION

In July of 1968 a milestone announcement was distributed internally to various codes in the Naval Oceanographic Office, advertising the "Newly Restructured NODC BT Data File." The announcement consisted of a computer generated listing with a printout of sample data contained in the new file (sample listing appended). The announcement on the title sheet referred to the use of "advanced free field and incremental formatting methods", and stated:

"The restructured file is contained in fixed length record blocks 4680 6-bit bytes long at 556 bytes per inch. Formerly contained on fourteen reels of tape in line image format (120 6-bit bytes per line image) with a blocking ratio of ten line images per tape block (1200 6-bit bytes), the BT file is now contained in 3900 restructured data blocks, on 1.25 reels of tape.

"The new format effects a reduction in computer memory storage requirements for the BT file in the ratio of about 9 to 1 and corresponding advantages in file accessioning and processing economy and speed."

Powerful formatting methods were used to provide the data compaction results cited in this announcement. The methods that were used are described briefly here, along with a detailed description of the new BT file format.

DATA COMPACTION METHODS

A. Regression Profile Substitution

The NODC BT format (refer to "Manual for Processing Bathythermograph Data," NODC Manual Series, National Oceanographic Data Center, Washington, D.C., 1964) encodes temperature data that have been visually digitized at five meter intervals from an ozalid copy of an analog profile trace on a bathythermograph slide. These digitized points, as many as sixty per BT profile, may be replaced by more economical sets of straight lines (Figure 1) selected to fit subsets of points that fall within a negligible distance of a straight line. ("Fitting Sets of Straight Lines to a Digital BT Profile", Yergen, Manuscript to be published.) A "three level tolerance algorithm" (Figure 2) was used to reduce NODC BT profile point sets to "regression point" sets, representing the union of lines fitted to straight line point subsets in the data. The regression point sets are recorded in the new file format.

A narrow thin line tolerance filter, ±0.08°Celsius, was used to process the NODC BT file. This value was selected to provide a noise-free recording of fine scale irregularities in a digital profile. It is estimated that an average of four

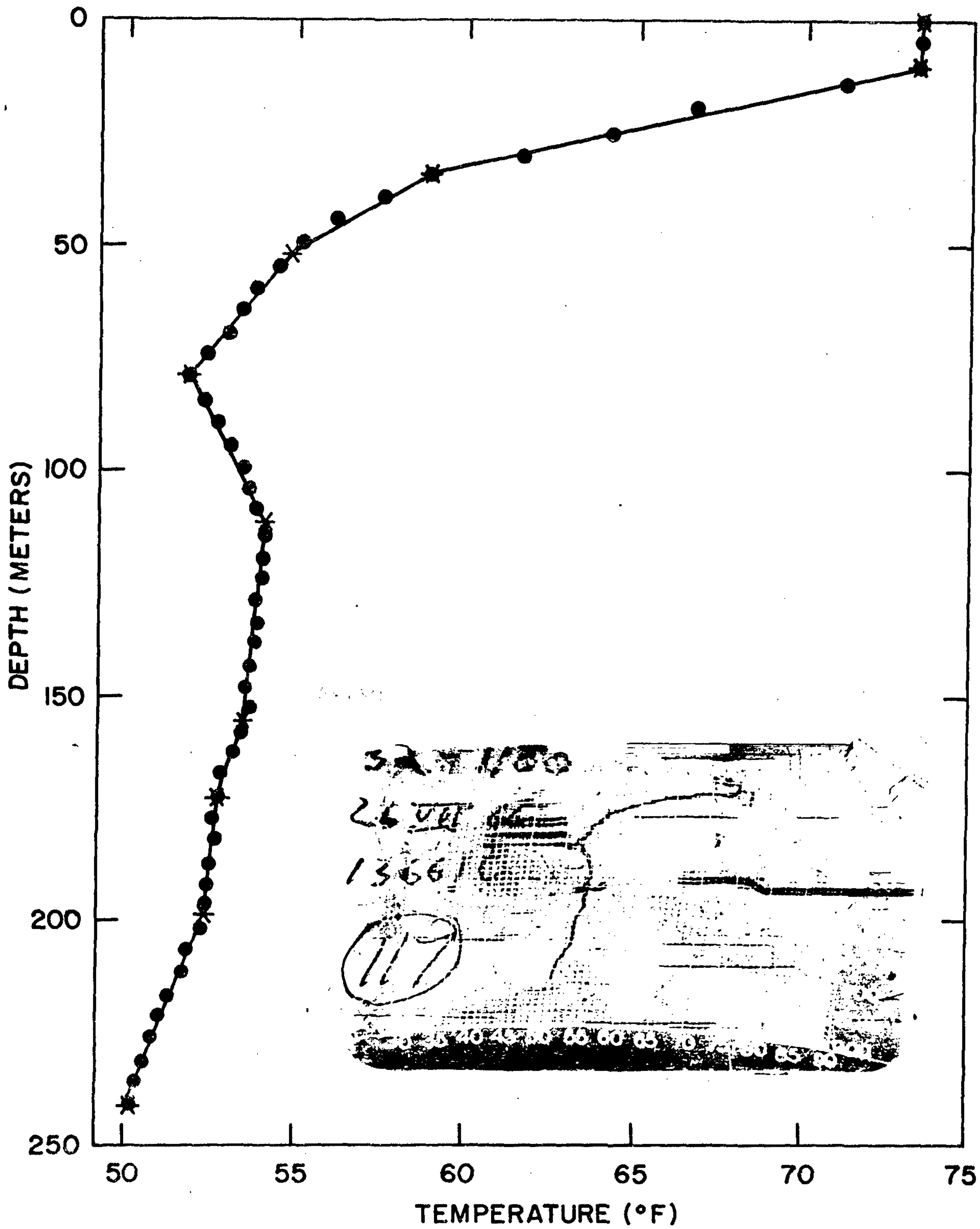

FIGURE 1. THE "SOLUTION PLANE" RECTILINEAR GRID IS DEFINED TO APPROXIMATE THE TEMPERATURE/DEPTH SCALE OF THE STANDARD CURVELINEAR GRID BATHYTHERMOGRAPH. THE BT PROFILE (INSET) WAS VISUALLY DIGITIZED AT FIVE METER OR FIFTEEN FOOT INTERVALS. DATA POINTS (DOTS) ARE PLOTTED ON THE SOLUTION PLANE. "REGRESSION POINTS" (ASTERISKS) WERE CALCULATED USING THE THREE LEVEL TOLERANCE ALGORITHM DESCRIBED IN THIS REPORT.

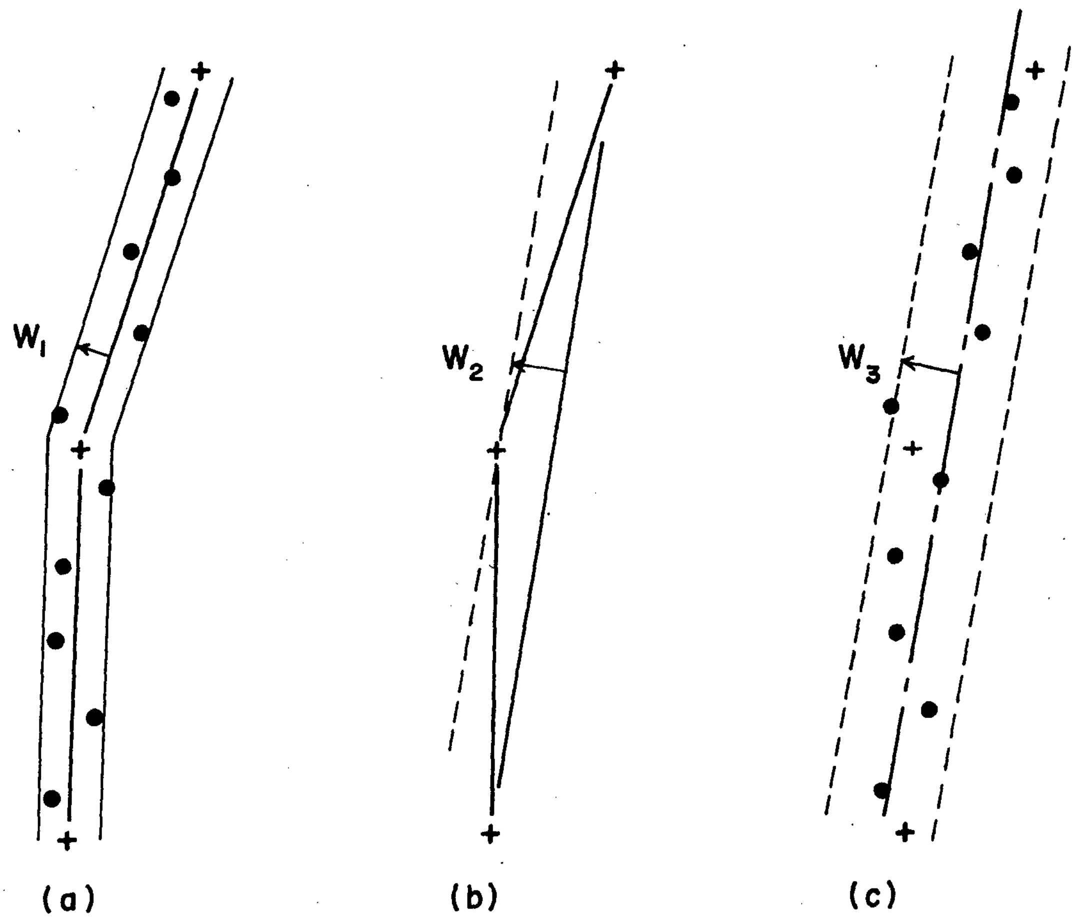

FIGURE 2. THE "THREE LEVEL TOLERANCE" ALGORITHM USES A FIRST LEVEL TOLERANCE VALUE, W_1, TO DETECT "THIN LINE SETS" IN A DIGITAL PROFILE (a). A SECOND LEVEL TOLERANCE VALUE, W_2, IS USED TO DETECT "VIRTUALLY COLINEAR" POINTS IN A REGRESSION SET (b). WHEN A VIRTUALLY COLINEAR POINT IS DETECTED A NEW REGRESSION SET IS COMPUTED, COMBINING LINE SETS ON EACH SIDE OF THE COLINEAR POINT. A THIRD TOLERANCE LEVEL VALUE, W_3, IS USED TO DETERMINE WHETHER THERE IS A DATA POINT THAT IS AN "INTOLERABLE" DISTANCE FROM THE RESULTING REGRESSION SET (c).

or five data points were reduced to one regression point in the product file using this tolerance value. It should be pointed out, however, that whereas sample depths were not recorded in the NODC BT format, (sample depths were implied by the position of sample temperatures in a fixed field profile format) the depth of a regression point is recorded in the new format. In addition, values which indicate the departures tolerated between NODC data temperatures and corresponding regression interpolated temperatures are recorded in the new file format. Hence, the ratio of profile memory reduction effected with the use of regression point profile representation is in the order of 4 or 5 to 3 rather than 4 or 5 to 1. It is estimated that this profile reduction provides an overall reduction ratio of about 4 to 3 in BT file memory requirements.

The principle advantage of regression profiles over fixed interval profiles is that they are "clean"; that is, the regression profiles are free of spurious irregularities which confound attempts to identify significant depths and gradients in the NODC format data. Hence, the regression data are easier to analyze, and may be more readily reduced to useful summaries with quick and simple programs. An additional advantage is that the regression profile is not format dependent upon the interval of digitization. Hence, expendable BT data, which are digitized at three meter intervals, may be processed for regression representation and merged with mechanical BT regression profiles, without altering the existing profile format or the interpretation of profile data.

B. Free Field Formatting

A powerful data compaction tool was effected with the use of free field formatting methods (refer to IR 68-81, "Free Field Formatting Scheme," Yergen, September 1968.) These methods were used to delete idle data fields from the new file. Thus, when a meteorological or sea state parameter is not reported in a BT record, the field needed to contain it is deleted from the record. Data fields are not deleted from records for which the corresponding data are available, however. This scheme provides potential advantages in the event that the BT file content is altered at some future time. For example, if surface reports are deleted from future BT data records the new file format will not have to be revised to contain the new BT data in minimum memory volume, nor will it be necessary to delete from future BT files the data that are currently carried in the file. Conversely, a variety of data items not currently reported in a BT record may be included in future records and merged with the current file without altering the existing file format.

Free field format technique further permits the use of a powerful compaction tool for the recording of digital profiles; namely, the "incremental formatting" technique. This technique consists of recording differences between successive

profile values rather than the values themselves. The technique is particularly powerful for compacting profiles which exhibit few and slight gradient irregularities. In an average BT profile, for example, the difference between temperatures at neighboring sample depths may be contained in a binary number field half as long as is needed to contain either temperature.

It is estimated that the file compaction ratio achieved with the use of free field formatting methods on the BT file is somewhat more than 3 to 2. This factor includes considerations of the data contained in the file to describe file format.

C. Binary Number Format

The free field formatting technique is used in combination with a binary number representation of file variables. It has been shown that the conversion from decimal in BCD format to binary yields a compaction ratio of about 2 to 1. (See IR 68-26, "A Guide to the Formatting of Oceanographic Data," Yergen, April 1968.) In fact, the free field formatting technique may be recommended for binary and not at all for decimal representation of file variables, considering the advantages of the binary number system for expressing data in minimum byte configurations. Thus, for example, just one bit is needed to indicate whether or not a data group is contained in a record. Presumably, six bits would be required to convey the same information in BCD format. The advantages of free field formatting in BCD would probably be erased by the cost of recording the format description data.

Hence, it might be estimated that the overall advantages of recording the BT file in a binary rather than BCD format is expressed in the file compaction ratio of about 6 or 7 to 2. This ratio includes the advantages of using free field formatting methods.

D. Additional Compaction

An estimated compaction ratio of about 4 to 3 was effected by deleting some verbose cruise identification data. A separate file of ship names, cruise consecutive numbers, country and institution identification codes, etc., may be maintained if desired. These data may be directly related to the BT data through NODC reference and consecutive numbers.

A file compaction ratio in excess of 3 to 2 was effected with the restructuring of the data format to variable length record blocks. Redundant signature data and inefficient extended overpunch designations carried in the NODC format were deleted from the restructured format, along with unused card columns.

The overall computer memory compaction ratio achieved for the file is about 9 or 10 to 1. This is the compaction ratio that is achieved for data that are stored on drum and disc or directly in processer memory. An additional compaction ratio of about 5 to 4 was achieved for tape by using a tape block length of 4680 6-bit bytes to record the data. This ratio was purchased by increasing the tape surface record to gap ratio from 3 to 1 (achieved with 1200 byte record blocks) to 11 to 1.

RESTRUCTURED BT FILE FORMAT

A. Standard Variable Length Record

A "record" is a computer memory block which contains data in a predetermined format. Format and block size are specified by the file originator. The new BT format conforms to a "standard" variable length record configuration. Other files that are contained in this configuration are the ocean station file, the sound velocity file, and a sample bathymetry file.

The standard variable length configuration provides a count of the 6-bit bytes, corresponding to the 6-bit column bytes on magnetic tape, containing a record. This count is recorded twice in the standard record configuration, once at the beginning and once for verification at the end of the record. In addition, a checksum byte is provided. The checksum is the sum obtained by adding the record bytes, excluding the checksum byte and the record counter bytes, modulo 2^6.

The record counter is contained in two 6-bit bytes, hence presupposes that the largest record to be contained in the standard format is less than 2^{12} or 4096 bytes in length. The count includes the five bytes that are used for recording checksum and count.

In summary, the standard variable length record configuration consists of the station counter, the users' data, the checksum, and the verifying station counter, in that order.

B. Record Identification Block

In addition, the BT format conforms to a "universal station format," which provides fixed field specifications for recording place and time of station. Other files that conform to this format are the ocean station and sound velocity files. It is anticipated that principle files (as opposed to derivative files, such as the sound velocity file) that conform to the universal station format will be merged with a "universal" archive file for possible "live atlas" use.

The universal station format is specified as follows:

ITEM	DESCRIPTION	BITS
1	Count of six bit bytes for this station	12
2	Marsden Square	10
3	Month	4
4	1° square	7
5	Year - 1850	9
6	Day of month	5
7	Time (hour)	5
8	Latitude hemisphere (0:north, 1:south)	1
9	Latitude degrees	7
10	Latitude minutes	6
11	Longitude hemisphere (0:west, 1:east)	1
12	Longitude degrees	8
13	Longitude minutes	6
14	File type	9

Additional data are appended to further identify a BT station as follows:

ITEM	DESCRIPTION	BITS
15	Time (minutes)	6
16	NODC reference number	17
17	NODC consecutive number	10
18	Quality code	4
19	Sample depth code	6
20	Depth interval count	6
21	Depth bit budget	4
22	Temperature bit budget	4
23	Surface water temperature	12

NOTES:

(The notations t(Q) and r(Q) are subsequently used to specify the truncated quotient and the remainder, respectively, of the rational fraction Q.)

1. File type is "4" for BT profiles in the format described by this paper.

2. "Quality code" specifies the quality description, Q, assigned to a BT profile, and the overpunch designation, X, which is used to indicate thick or barely legible grid lines. If X=0 for no overpunch and 1 for overpunch,

$$Q = t(CODE/2) + 1, \quad X = r(CODE/2)$$

(Q is assigned values 1 through 5).

3. Maximum profile depth, MD, is calculated from the sample depth code using the formula,

$$MD = 5(CODE - 1) \text{ meters.}$$

4. Depth interval count is a count of the regression points used to code the profile, less one.

5. Bit budgets, items 21 and 22, provide a count of the bits used to code the differences between successive regression points.

6. Surface depth is assumed to be zero. Surface temperature is coded in an eleven bit field, assuming four radix bits, followed by a sign designator bit (0:positive, 1:negative).

C. Profile Code Block

Differences between depths and temperatures at successive regression points are coded in depth-temperature groups, in bytes of budget specified length (identification block items 21 and 22). Depth and temperature values at successive regression points may be derived by arithmetically accumulating successive differences, using surface values (zero for depth, and item 23 for temperature) as initial accumulations. Depth differences are coded in meters. Temperature differences are coded in sixteenths of a degree Celsius (four radix bits) followed by a sign designator bit (0:positive, 1:negative). The temperature bit budget count includes the sign bit.

Depth and temperature unit intervals were chosen to represent approximately equivalent intervals on the standard scale mechanical BT grid. The temperature unit interval of 1/16 has the advantage that it is easily manipulated in the binary base arithmetic of most computer processers. Temperature and depth each are recorded to excessive precision (see IR 68-26) to satisfy the instincts of conservative file users. Hence, this file is designed as an "archive file", and not as a special purpose file that has been streamlined for live atlas use.

D. Surface Report Block

The profile code block is followed by a surface report flag bit. When the flag bit is set to zero, the surface report block is not recorded, and the RMS code block follows the flag. When the flag bit is set to 1, the surface report block is recorded following the flag.

Each item in the surface report block is identified with a flag bit. If the flag is set to zero then the item has not been recorded. If the flag is set to 1, then the item is reported in a byte of the specified size following the flag bit. The order of surface report data and byte sizes are specified in the following table.

ITEM	DESCRIPTION	BITS
1	Bottom depth, meters	14
2	Weather code	4
3	Wave period	4
4	Pressure code (P-945 millibars)	7
5	Dry bulb temperature	10
6	Wet bulb temperature	10
7	Wind code	8
8	Cloud code	7

NOTES:

1. WMO weather code = recorded code - 1.

2. Wave period = CODE + 1 seconds; equals or exceeds 16 if CODE = 15.

3. Pressure = CODE + 945 millibars.

4. Dry bulb temperature = (CODE-500)/10, degrees Celsius.

5. Wet bulb temperature = (CODE-550)/10, degrees Celsius.

6. Wind code is composite of speed, CV, and direction, CD, codes:

$$CV = t(CODE/10), \quad CD = r(CODE/10).$$

The wind speed code is related to Beaufort scale code, CB, by the relation $CB = t(CV/2)$, and is related to speeds expressed in knots, KV, by the approximations:

(a) $$KV = .2125 + (CV-1.)\left[.55 + .1375(CV-1.)\right],$$ when $1 \leqslant CV \leqslant 9$.

(b) $$KV = 1.8929 + (CV-1.)\left[1.4762 + .05952(CV-1.)\right],$$ when $9 < CV \leqslant 23$.

(c) wind speed is not reported when $CV = 0$.

(d) wind is hurricane force and is otherwise unspecified when $CV = 24$.

The direction code, CD, is interpretted as follows:

CD	DESCRIPTION	CD	DESCRIPTION
0	Not reported	5	South
1	North	6	Southwest
2	Northeast	7	West
3	East	8	Northwest
4	Southeast	9	Variable

7. When cloud code = 127, sky is obscured; otherwise cloud code is a composite of cloud amount code CA, and cloud type code CT.

$$CA = t(CODE/11),$$

$$CT = r(CODE/11).$$

(a) When CA = 0, cloud amount is not reported.

Otherwise, WMO cloud amount code = CA - 1.

(b) When CT = 0, cloud type code is not reported.

Otherwise, WMO cloud type code = CT - 1.

E. RMS Code Block

The root mean square values of differences between NODC format digital BT temperature data and interpolated temperature values in a straight line regression interval are recorded for each interval. The values are recorded in Celsius degrees, to four radix bits.

The first item of data in the RMS code block is the RMS bit budget, which specifies the number of bits used to encode individual RMS values. The bit budget is contained in a 4-bit byte. This byte is followed by RMS values equal in number to the number of profile intervals. The RMS values, and not the differences between successive RMS values, are coded.

ADDITIONAL COMMENTS

A. Format Modifications

This file is supposed to be an "archive file". It may be modified in several ways to provide a "climatological summary file". For example, permissive tolerance filters may be used to smooth short interval, short amplitude variations from the profiles, so that a profile may be represented with fewer regression points. This presupposes that short interval variations in a profile trace represent short term variations that may be regarded as "climatologically insignificant". At the same time, individual points may be recorded in smaller bytes by eliminating insignificant bits. Moreover, BT profiles of low quality may be weeded from portions of the file that provide a high density of BT data in the time and space.

A number of the changes that may apply to the derivation of a climatological or summary file can only evolve through feed back experimentation, best performed using live atlas analytical methods. In my opinion it is a waste of time to speculate at any length on the content and description of data summary bases to be used in a live atlas configuration, until the live atlas hardware environment exists to provide a practical performance rating of alternative summary schemes.

B. File Access

At the present time, the variable length BT stations are recorded in fixed length tape blocks 4680 6-bit bytes long. This configuration may be changed to take advantage of system software sort/merge programs. Considering the complexity of the BT file format and the possible variations that it may undergo, it might seem that the file is inaccessible to the average potential file user.

However, standard accessioning subprograms have been developed along with the file, to provide reasonably legible data listings as a control of file format quality. These subprograms deblock the tape records, unpack the BT profiles and edit them to a common program memory block, in a form that can be used to read data for significant portions of one BT record at a time using a fortran "CALL" statement.

The accessioning subprograms will be modified to accomodate any changes effected in the BT file format structure. The subprogram product, consisting of BT data in the predetermined memory format, will remain constant however. Hence a user of the accessioning program may assume that the data format will remain constant, and that his programs will not require changes to accomodate experimental changes in the BT file format. The same user coded program may be used to analyze BT data whether in archive or climatological format.

A sample listing of BT data retrieved from the restructured file using the BT data accessioning program is shown in Appendix A.

BT PROFILES NODC DATA * YERGENS FORMAT *

MSQR DEG MO

042 47 APR 19 1961 TIME 0300

REF 04156 14 37 N 057 59 W
CSC 887 QUAL 2

BOTTOM DEPTH 4572 M
WIND 18 KNOTS, E
WAVES 3
TD=26.7 TW=23.3
P=1015 MB
CUMULUS COVER 2

	DEPTH M	TEMP C	GRAD C/M	RMS C
1	0	26.9		
2	19	26.7	-0.01	0.06
3	43	26.3	-0.02	0.06
4	81	26.1	-0.00	0.06
5	97	26.4	0.02	0.06
6	118	24.9	-0.07	0.13
7	128	23.9	-0.09	0.13
8	131	23.8	-0.06	0.06
9	166	20.9	-0.08	0.13
10	177	19.8	-0.10	0.06
11	220	17.6	-0.05	0.06
12	240	17.1	-0.03	0.06

042 47 APR 30 1966 TIME 2000

REF 07126 14 09 N 057 53 W
CSC 335 QUAL 2

BOTTOM DEPTH 4206 M
WIND 3 KNOTS, NE
WAVES 1
CUMULUS COVER 2

	DEPTH M	TEMP C	GRAD C/M	RMS C
1	0	26.8		
2	27	26.7	-0.00	0.06
3	43	25.9	-0.05	0.06
4	95	24.9	-0.02	0.06
5	137	20.4	-0.11	0.06
6	158	18.9	-0.07	0.13
7	167	18.5	-0.05	0.06
8	210	17.1	-0.03	0.06

042 47 MAY 18 1959 TIME 2005

REF 06132 14 16 N 057 19 W
CSC 499 QUAL 2

BOTTOM DEPTH 5029 M

	DEPTH M	TEMP C	GRAD C/M	RMS C
1	0	27.3		
2	36	26.8	-0.01	0.06
3	71	25.7	-0.03	0.06
4	84	25.5	-0.01	0.06
5	100	24.9	-0.04	0.06
6	121	23.5	-0.07	0.06
7	162	18.7	-0.12	0.13
8	180	18.4	-0.02	0.06
9	185	17.4	-0.20	0.00
10	215	16.0	-0.05	0.06

APPENDIX A

APPENDIX: Sample listing of BT profiles, retrieved from the restructured file. Regression points are listed, along with the gradients and RMS departures for successive intervals between regression points.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* U. S. NAVAL OCEANOGRAPHIC OFFICE	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

A FREE FIELD RAPID ACCESS FORMAT FOR BT PROFILES

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Informal Report

5. AUTHOR(S) *(First name, middle initial, last name)*

Walter E. Yergen

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
March 1969	15	0

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S) IR NO. 69-40
b. PROJECT NO. HF 05-243-301 c. 707-FC-DAR d.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY U. S. Naval Oceanographic Office

13. ABSTRACT

A format of the restructured bathythermograph file is specified. The methods used to achieve a file memory requirement reduction ratio of about 9 to 1 are described.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
A FREE FIELD RAPID ACCESS FORMAT FOR BT PROFILES						